Early Journal Content on JSTOR, Free to Anyone in the World

This article is one of nearly 500,000 scholarly works digitized and made freely available to everyone in the world by JSTOR.

Known as the Early Journal Content, this set of works include research articles, news, letters, and other writings published in more than 200 of the oldest leading academic journals. The works date from the mid-seventeenth to the early twentieth centuries.

We encourage people to read and share the Early Journal Content openly and to tell others that this resource exists. People may post this content online or redistribute in any way for non-commercial purposes.

Read more about Early Journal Content at http://about.jstor.org/participate-jstor/individuals/early-journal-content.

JSTOR is a digital library of academic journals, books, and primary source objects. JSTOR helps people discover, use, and build upon a wide range of content through a powerful research and teaching platform, and preserves this content for future generations. JSTOR is part of ITHAKA, a not-for-profit organization that also includes Ithaka S+R and Portico. For more information about JSTOR, please contact support@jstor.org.

ON A SHORT METHOD OF LEAST SQUARES.

BY BURTON H. CAMP.

1. **Introduction.** In one of the "Scientific Papers of the Bureau of Standards"* it is shown that, when the independent variables of a set of observational equations are equi-spaced, certain simplifications are available which greatly shorten the least square solution. These simplifications are of the same character as those which have been used by statisticians in computing the coefficients of regression,† and the trend in a time series.‡ They may be described in a particular case as follows:

Consider the type equation:

$$Z = A + BX + CY, \tag{1}$$

where X and Y are independent variables, giving n equations of condition to determine A, B and C. To obtain the first normal equation, add the equations of condition. To obtain the second, arrange them in order of ascending X's, multiply the successive equations through by the successive terms of any ascending, equi-spaced sequence of integers, and then add the results. To obtain the third normal equation, arrange in order of ascending Y's, multiply by a sequence of integers, as before, and add. It is sometimes supposed necessary that the sequence of integers shall be special in form, but this is in fact immaterial, all ascending, equi-spaced sequences leading to the same results. It will usually be convenient, however, to choose the sequence:

$$-n+1,\ -n+3,\ \ldots,\ -1,\ +1,\ \ldots,\ +n-1, \tag{2}$$

when n is even; and the sequence:

$$-(n-1)/2,\ -(n-1)/2+1,\ \ldots,\ 0,\ \ldots,\ (n-1)/2, \tag{3}$$

---

* No. 388, Adjustment of parabolic and linear curves to observations taken at equal intervals of the independent variable, by Harry M. Roeser, July, 1920, Department of Commerce, U. S. A.

† E. g., Yule, Theory of statistics, Griffin and Co., Ltd., 1911, pp. 181 ff.

‡ E. g., Persons, The Review of economic statistics, Preliminary Volume I, 1919, p. 13, foot-note 1.

when n is odd. These sequences will be assumed in the calculations to follow and in the resulting expressions (10) for the errors. It will be shown that, if the X's and the Y's are not equi-spaced but are approximately so, this short method may be used with negligible error, and expressions for the error will be found. It often happens that only approximate equi-spacing is possible. One may have only partial control of the independent variables, X, Y, and it would be very difficult or impossible to choose them exactly equi-spaced, as, for example, when selecting a star list for time observations with a transit. Sometimes, when one does have apparent control of these variables and has set his instrument so that they are equi-spaced, the spacing is disturbed after allowance is made for instrumental errors; the same effect is produced if the observations have slightly varying weights. It is, therefore, desirable that an expression for the error in using the short method should be obtained, and it should be in such a form that it may be computed quickly, after the short determination is made. Further, it should be possible to estimate roughly the size of the error to be expected, before *any* computation has been made.

If the order of the X's is the same as the order of the Y's the method as outlined above fails. An important case where this happens is when $Y = X^2$ in the type equation. It is feasible to make special arrangements in such cases, but they will not be discussed here. The results in the special cases, A, B, or C equal to zero, will be given, however. The computer may use the summary in § 3 below without reading the theory in § 2.

2. **Theory.** Let i be the representative integer of the sequence (2), or (3), while the second normal equation is being formed, and j the representative integer for the third normal equation. Let X_0 denote the middle X, when n is odd, the point half way between the two middle X's, when n is even. Let h denote the "normal" space interval of the X's, this to mean that value which will make the sum of the squares of the errors (e) made in writing, $X - X_0 = hi$, a minimum. For the Y's, let Y_0, k, and d be the analogous letters. Let r denote the residual, $Z - (A + BX + CY)$. Thus,

$$(4) \qquad h = \frac{\sum i(X - X_0)}{\sum i^2}, \qquad k = \frac{\sum j(Y - Y_0)}{\sum j^2}.$$

The true normal equations, resulting from an exact application of the least square method may be written

$$(5) \quad \sum r = 0, \quad \sum (hi + e + X_0)\, r = 0, \quad \sum (kj + d + Y_0)\, r = 0,$$

and, by virtue of the first of these equations, the $\sum X_0 r$ and $\sum Y_0 r$ of the second and third drop out. Make the transformations

$$(6)\quad X = x + \overline{X},\ Y = y + \overline{Y},\ Z = z + \overline{Z},\ \text{where } \sum x = \sum y = \sum z = 0.$$

Equations (5) become, after reduction,

$$(7)\quad \begin{aligned} \overline{Z} &= A + B\overline{X} + C\overline{Z}, \\ \sum hiz &= B\sum hix + C\sum hiy - \sum er, \\ \sum kjz &= B\sum kjx + C\sum kjy - \sum dr. \end{aligned}$$

Solving by determinants the last two of (7), one gets the true values of B and C; putting $e = d = 0$ in the results gives the values of B and C that would be obtained by the approximate method under consideration. The differences will be denoted by dB and dC, and are

$$(8)\quad \begin{gathered} dB = \frac{1}{\Delta}\begin{vmatrix} \sum er & h\sum iy \\ \sum dr & k\sum jy \end{vmatrix}, \qquad dC = \frac{1}{\Delta}\begin{vmatrix} h\sum ix & \sum er, \\ k\sum jx & \sum dr, \end{vmatrix}, \\ \Delta = hk\begin{vmatrix} \sum ix & \sum iy \\ \sum jx & \sum jy \end{vmatrix}. \end{gathered}$$

It is now necessary to find approximate values of these expressions in terms of quantities which will be found in the course of the short method computation, or of other known quantities.

Let s_A, s_B, s_C refer to the mean square deviations*, called by the statistician standard deviations, of A, B, C, respectively, due to errors of sampling. They may be found approximately by using the values of B and C obtained by the approximate method. Thus, by the ordinary theory of errors,

$$\begin{aligned} s_B^2 &= \frac{h^2k^2}{\Delta^2}\left[(s_r^2\sum i^2)\ (\sum jy)^2 + (s_r^2\sum j^2)\ (\sum iy)^2\right] \\ &= \frac{h^2k^2s_r^2}{\Delta^2}\left[(\sum jy)^2 + (\sum iy)^2\right]\sum i^2, \end{aligned}$$

where $s_r = \sqrt{\frac{1}{n}\sum r^2}$, and this also equals the mean square deviation of the z's, about the mean plane. Solving this equation for Δ^2 and substituting in (8), one gets

---

* One might as well use "probable" deviations, or probable errors, by multiplying the standard deviations by the appropriate constants.

$$(9)\qquad dB = \frac{\begin{vmatrix} \sum er & h\sum iy \\ \sum dr & k\sum jy \end{vmatrix}}{hks_r\sqrt{(\sum jy)^2 + (\sum iy)^2}} \cdot \frac{s_B}{\sqrt{\sum i^2}}.$$

Now e and r are uncorrelated, and so the most probable value of their coefficient of correlation, $\sum er/s_e s_r n$, where $s_e = \sqrt{\sum e^2/n}$, is zero, and it is almost certain* that this quantity does not depart from zero by more than three times its own standard deviation; it is probable that this departure is, in fact, less than the standard deviation. Thus, almost certainly,*

$$\left|\frac{\sum er}{n\, s_e\, s_r}\right| < \frac{3}{\sqrt{n}},$$

and therefore $\left|\sum er\right| < 3\, s_r\, s_e \sqrt{n}$; and similarly $\left|\sum dr\right| < 3\, s_r\, s_d \sqrt{n}$. On substitution in (9), it is found that, almost certainly,†

$$|dB| < 3s_B\sqrt{\frac{n}{\sum i^2}} \cdot \frac{s_e\left|\sum jy\right| k + s_d\left|\sum iy\right| h}{hk\sqrt{(\sum jy)^2 + (\sum iy)^2}}.$$

An approximation to this expression may be made by putting $y = Y - \bar{Y}_0$, since $\bar{Y} - \bar{Y}_0$ is small. The result may then be simplified by using the fact that $\sum i = \sum j = 0$, and inserting the values of h and k from (4). Finally, then,

$$|dB| < 3s_B\sqrt{n\sum i^2}\; \frac{\dfrac{\left|\sum jY\right|}{\left|\sum iX\right|}s_e + \dfrac{\left|\sum iY\right|}{\left|\sum jY\right|}s_d}{\sqrt{(\sum jY)^2 + (\sum iY)^2}},$$

$$(10)\dagger\qquad |dC| < 3s_B\sqrt{n\sum i^2}\; \frac{\dfrac{\left|\sum iX\right|}{\left|\sum jY\right|}s_e + \dfrac{\left|\sum jX\right|}{\left|\sum iX\right|}s_d}{\sqrt{(\sum iX)^2 + (\sum jX)^2}},$$

$$dA = -\frac{1}{n}\left(dB\sum X + dC\sum Y\right).$$

---

* On the assumption of a Gaussian distribution, the probability that the coefficient of correlation will exceed its standard deviation is 0.32, and the probability that it will exceed three times its standard deviation is only 0.0027.

† See preceding foot-note. In the examples following the theory, limits obtained from (10) are described as "nearly certain". Limits obtained from (10) without the factor 3 are described as "probable".

These formulae hold good only if the i and j sequences are chosen as in (2) and (3). All the sums involved occur in the normal equations as determined by the short method. After these sums have been found, a slide rule computation suffices for dB/s_B, and dC/s_C. It is these ratios, and not the absolute values of the errors which are important in determining the question whether the short method is justified. For example, if $\lambda = dB/s_B$ is as small as 0.01, then there is a probability of 0.996 that dB, the error introduced by the short method, is numerically less than the necessary error in B, due to the accidental errors of observation. The corresponding probabilities for other λ's are given by a Gaussian table, as follows:

λ	prob.	λ	prob.	λ	prob.
.005	.996	.10	.92	.4	.69
.01	.992	.20	.84	.5	.62
.05	.957	.30	.76	1.0	.32

Similar remarks apply to C. As will become obvious from the examples to follow, it is necessary to have only an approximate estimate of the sizes of s_e and s_d. These are sometimes obvious from the nature of the problem (e. g., Ex. 2 below). In other cases they must be computed. The labor of doing this is slight. Consider, for example, s_e. The values of $\sum iX$ and $\sum i^2$ from which h is found have already been obtained in the short solution. Then $e_i = X_i - X_0 - hi$, $s_e^2 = \sum e^2/n$. It follows from the last equation of (10) that dA may be made zero, *whether the X's and Y's are nearly equi-spaced or not,* provided they are so chosen that $\sum X = \sum Y = 0$. This is a consideration worth keeping in mind if A is the quantity most desired from the observations. It follows also that, if $\sum X/n$ and $\sum Y/n$ are small, the approximate solution for A may be used, instead of the long solution, with small error. It seems to the author that this method could be used to advantage in getting the time from transit observations. Tables already in use in connection with other methods make it easy to select the stars properly (Cf. Ex. 3 below).

The expressions just found are available after the approximate normal equations have been found. It is now desirable to shorten them by further approximations so that they may be made to yield good estimates of the errors, in advance of *any* computation. It will be necessary to assume now that the order of the X's is uncorrelated with the order of the Y's. This is frequently the case, but there are important instances in which it is not, and in such instances the subsequent approximations must not be used. Approximately,

$$(11)\ \sum i(Y-Y_0)=\sum j(X-X_0)=0,\ X_i-X_0=ih,\ Y_j-Y_0=jk.$$

Omitting the coefficients 3 in (10), and substituting (11), one obtains, as a provisional value for dB,

$$(12)\qquad dB=\frac{s_B\,s_e}{h}\sqrt{\frac{n}{\sum i^2}}.$$

If n is odd, $\sum i^2=n(n-1)(n+1)/12$. If n is even, $\sum i^2=n(n-1)(n+1)/6$ but in the latter case h and k are only half what they are in the former case, to that, if H and K be approximately the average intervals between the X's, and between the Y's, respectively, then in all cases:

$$(13)^*\ \frac{|dB|}{s_B}=\frac{s_e}{H}\cdot\frac{2\sqrt{3}}{\sqrt{(n+1)(n-1)}},\qquad \frac{|dC|}{s_C}=\frac{s_d}{K}\cdot\frac{2\sqrt{3}}{\sqrt{(n+1)(n-1)}}.$$

The quantities s_e/H, s_d/K, measure the divergences from equi-spacing, and it is natural than they should appear as factors in the expressions for the errors. In most problems where any real attempt is made at equi-spacing it is possible to make s_e/H and s_d/K as small as $1/3$. Indeed, ten times this smallness is to be expected frequently. In very bad cases, as in Example 3, where the conditions of the problem are such as to interfere with equi-spacing, these quantities may be as large as 1.5. It follows from (13) that dB/s_B and dC/s_C approach zero with $1/n$. One must compute dA from the formula of (10).

3. **Summary.** Let dA, dB, dC be the errors made in A, B, C if the short method of § 1 is used in connection with (1) instead of the exact least square solution. To find dB/s_B and dC/s_C, provisionally, substitute in (13), s_B and s_C being the mean square deviations† of B and C due to the fluctuations of the observations themselves. The values of s_e/H and s_d/K may be taken as stated under (13). To find dA/s_A, then use the last equation of (10). Formulae (13) are independent of the particular type of ascending equi-spaced sequence of integers used as multipliers, but formulae (10) are not.

To find closer approximations to these relative errors, after the short computation has been made, use equations (10), noting what is said in the remainder of the paragraph containing (10). $\sum i^2$ is best found from a short table which the computer may as well make for himself, although several such tables are published for large values of n‡.

---

* Probable values. Insert the factors 3 to obtain more certain limits. See preceding foot-note.

† Also called "standard deviations".

‡ E. g., "Tables for statisticians and biometricians", by Karl Pearson, Cambridge University Press, 1914.

For the special case $Z = BX + CY$, the same formulae may be used if, and only if, X_0 and Y_0 are numerically small, relative to those values of X and Y which are distant from X_0 and Y_0. Otherwise, it is not recommended that the short method be used.

For the special case $Z = A + BX$, formulae (10) and (13) take on the simple forms:

$$(10a)^* \qquad \frac{|dB|}{s_B} = \sqrt{n\sum i^2} \cdot \frac{s_e}{\left|\sum iX\right|}, \qquad dA = -dB\frac{\sum X}{n},$$

$$(13a)^* \qquad \frac{|dB|}{s_B} = \frac{s_e}{H} \cdot \frac{2\sqrt{3}}{\sqrt{(n+1)(n-1)}},$$

and the various remarks made in connection with (10) and (13) apply of course to these special forms.

4. **Examples.** *Example 1.* For the first illustration a rather bad case will be chosen, a group of five stars composing a time set, with only the crudest attempt at equi-spacing. They are chosen from the unweighted set in Campbell's "Elements of Practical Astronomy", page 150, and in the following observational equations, the letters A, B, and C, correspond with Campbell's x, c, and a, respectively. The equations are

$$\begin{aligned}
-1.66 &= A - 3.813\ B + 3.395\ C \\
-0.13 &= A + 1.007\ B + 0.587\ C \\
+0.98 &= A + 2.915\ B - 1.353\ C \\
-0.02 &= A - 1.714\ B - 0.357\ C \\
+0.47 &= A - 4.898\ B - 2.875\ C
\end{aligned}$$

The results, as obtained by the exact method of least squares and also by the approximate method are:

	True Values	Approximate Values	Differences, called "errors" above	Standard Deviations
A	.0488	.0512	.0024	.024
B	.1266	.1289	.0023	.008
C	—.3658	—.3689	.0030	.011

---

* Probable values. See foot-notes to (10) and (13).

The next table gives the provisional and final estimates of the maximum values of the errors, as derived from the formulae of this paper, also their true values. In using equation (13) it was assumed that $s_e/H = s_d/K = 1/3$. It turned out that these numbers were 0.27 and 0.36, respectively. From the table in § 2, it follows that, assuming a normal distribution of error, the fluctuation of A due to the observations themselves, would, in 92 cases out of 100, exceed the difference (.0024) introduced by the short method of computation, and the corresponding numbers for B and C are 77 and 78 cases. It would seem to the author that, in examples like this, the short method is usually justifiable, but that it might not be sometimes. It is, however, not the object to present an argument for the short method, but to present a criterion by which the error to be committed by it may be estimated.

	Limits* by (13)		Limits by (10)		True Values
	Probable	Nearly Certain	Probable	Nearly Certain	
$\lvert dA/s_A \rvert$	—	—	—	—	.104
$\lvert dB/s_B \rvert$	.24	.72	.20	.60	.287
$\lvert dC/s_C \rvert$	.24	.72	.26	.75	.282

Example 2. A very good case will now be considered, a set of observations used by Roeser† with the equi-spacing disturbed by the arbitrary addition of small quantities. These small quantities were chosen at random from a table of numbers subject to the limitation that they should lie between .05 and —.05. The equations and tables following have interpretations similar to those of Example 1. It is obvious from a mere glance at the equations that s_e/H is not greater than about .004, and this was the value used in the provisional estimate. The true value is 0.0028. If this had been used instead of 0.004, the provisional estimate for dB/s_B would have differed from the true value by only about 4 per cent., thus illustrating the closeness of the inequality. It is not desirable, however, in practice to compute s_e/H in advance, and accordingly the rough estimate (0.004) is used. With this rough approximation, formula (13) shows that the value of dB/s_B will be so small as to make the difference between the two methods entirely negligible. This is exhibited in the table below.

---

* The first column is obtained from (13). The second column is three times the first. See also foot-note to (10).

† Loc. cit., p. 371. See also Weld, Theory of errors and least squares, 1916, p. 76.

Conditional Equations

$$\begin{aligned} 0.116 &= A + 10.05\,B, & 0.595 &= A + 60.01\,B,\\ 0.205 &= A + 19.95\,B, & 0.675 &= A + 69.98\,B,\\ 0.295 &= A + 29.99\,B, & 0.760 &= A + 79.93\,B,\\ 0.388 &= A + 40.01\,B, & 0.850 &= A + 90.04\,B,\\ 0.503 &= A + 49.97\,B, & 0.926 &= A + 100.01\,B. \end{aligned}$$

	True	Approximate	Difference	Standard Deviation
A	.028478	.028425	.000053	.00315
B	.0091417	.0091418	.0000001	.00011

	Limits by (13a)		Limits by (10a)		True
	Probable	Nearly Certain	Probable	Nearly Certain	
dA/s_A	—	—	—	—	.00168
dB/s_B	.00139	.00417	.00099	.00297	.00096

A is not so well determined as B. This is because the origin is far from the middle X.

Example 3. Finally, a very bad case will be considered, viz., Campbell's eleven weighted equations* out of which the data of Example 1 were selected. Here the method of choice of stars was such as deliberately to prevent equi-spacing, as a rough plotting of U, X, and Y would show. The typical equation is

$$Z = AU + BX + CY,$$

a type not considered in the theory, but formulae (13) are valid. In such bad cases as this, the divergence from equi-spacing is so great that the ratios like s_e/H, which measure it, cannot be assumed small, and so it is advisable to estimate them by a graph. If, for example, X_i be plotted as a function of i, it will be seen that s_e is about unity and H about 2/3, making s_e/H about 1.5. The corresponding fractions for U and Y may be found in a similar manner to be about 1.2 and 1.0, respectively.

---

* Loc. cit., p. 155.

	True	Approximate	Difference	Standard Deviation		Limits by (13)		True
						Probable	Nearly Certain	
A	.053	.053	.000	.017	dA/s_A	.39	1.2	.01
B	.115	.122	.007	.012	dB/s_B	.47	1.3	.58
C	—.383	—.386	.003	.020	dC/s_C	.32	1.0	.15

Thus, even in this case, the provisional formula (13) is adequate to show that the errors are probably as small as half the standard deviations of the quantities measured. The very small actual error in the time correction, A, is due in part to the fact that the means of the X's and Y's are small. This example, taken together with Example 1, also serves to illustrate how the disturbance due to bad spacing can be masked by an increase in the number of equations of condition. This result is to be expected from the form of (13), as was noted in the sentence preceding the last of § 2. The value of the short method in saving labor increases greatly with the number of equations of condition, and it is therefore particularly comforting to know that at the same time the error in using it diminishes.

5. **Conclusion.** The short method works sufficiently well to justify its use, certainly in cases as good as Example 2, probably in cases like Example 1, and perhaps (depending on the accuracy desired) in cases as bad as Example 3. In any case the goodness of the method may, with sufficient accuracy, be judged in advance by the criteria (13) or (13a). When the short method is used, the gain in time is very great, especially if one employs a computing machine. The author used an electrically driven listing machine, but an ordinary adding machine would have done almost as well. All the twelve coefficients of the normal equations in Example 3 for the eleven equations of condition were found in this way in less than twenty minutes. In the short method, all the multiplications necessary for the coefficients like $\sum iX$ are by small integers. Thus, with an adding machine, the products, iX, may be found and added at the same time.

WESLEYAN UNIVERSITY,
May, 1922.